Deep-learning for dysgraphia detection in children handwritings

Andrea Gemelli*
andrea.gemelli@unifi.it
Università di Firenze
Florence, Italy

Simone Marinai
andrea.gemelli@unifi.it
Università di Firenze
Florence, Italy

Emanuele Vivoli*
emanuale.vivoli@unifi.it
Università di Firenze
Florence, Italy

Tamara Zappaterra
tamara.zappaterra@unife.it
Università di Ferrara
Ferrara, Italy

ABSTRACT

Early identification of dysgraphia in children is crucial for timely intervention and support. Traditional methods, such as the Brave Handwriting Kinder (BHK) test, which relies on manual scoring of handwritten sentences, are both time-consuming and subjective posing challenges in accurate and efficient diagnosis. In this paper, an approach for dysgraphia detection by leveraging smart pens and deep learning techniques is proposed, automatically extracting visual features from children's handwriting samples. To validate the solution, samples of children handwritings have been gathered and several interviews with domain experts have been conducted. The approach has been compared with an algorithmic version of the BHK test and with several elementary school teachers' interviews.

CCS CONCEPTS

• **Applied computing → Online handwriting recognition**.

KEYWORDS

dysgraphia detection, handwriting recognition, computer vision

ACM Reference Format:
Andrea Gemelli, Simone Marinai, Emanuele Vivoli, and Tamara Zappaterra. 2023. Deep-learning for dysgraphia detection in children handwritings. In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3573128.3609351

1 INTRODUCTION

Specific Learning Disorder (SLD) is a broad term that encompasses difficulties in children's ability to learn and produce written language and to process information in areas such as reading, writing, and mathematics. Dysgraphia is a type of SLD that specifically affects writing abilities. Children with dysgraphia often struggle with fine motor control, letter formation, and spacing, leading to difficulties in producing legible and coherent written work [6, 12]. The prevalence of SLD and dysgraphia is estimated to be between 10-30% of the population [10] and it is a common cause of academic underachievement in school-aged children [2, 14]. Despite its high prevalence and significant impact on children's learning and development, there is no universally recognized definition or diagnosis of dysgraphia, and different diagnostic tests are employed in different countries. This is partly due to the complex and multifaceted nature of the disorder and to the difficulty in distinguishing it from other related conditions such as developmental coordination disorder [3] and attention deficit hyperactivity disorder [1]. As a result, it has become critical in the last years to propose automatic and innovative approaches to assist experts in their diagnosis and to provide screening tools for early identification. The detection of dysgraphia is a well-known task in handwriting recognition, where scholars have proposed several solutions, ranging from algorithmic versions of dysgraphia assessments methods, to machine learning based techniques. Standardized tests, such as the Brave Handwriting Kinder (BHK) test which utilize an evaluation grid for the morphological graphology of children, can be employed to assess the morphological quality of signs and the spacing of graphemes (an example of children's handwriting for the Italian BHK test is shown in Fig. 1). Some of the criteria considered are the size of the writing, the misaligned left margin, and the shifting writing line. Among others, two algorithmic versions of BHK where proposed to analyze handwriting images and to extract scores automatically: a shorter but effective screening tool based on BHK, called SOS [15] and a software called TestGraphia [4]. Inspired by these previous works, a custom BHK algorithmic version has been developed in this research (described in Section 2) as a baseline for comparison. Only offline document images are considered in [4], thus missing all the information related to online handwriting (such as speed, stroke order, pressure, etc.), as previously demonstrated to be effective using a commercial digitalization tablet [7, 13]. Taking inspiration from machine learning statistical methods, in [5?] the authors adopted hand-crafted features that require specific feature extraction algorithms. For instance, in [5], suitable features were selected using weighted k-nearest neighbor and employed as input for an AdaBoost classifier, resulting in an accuracy of 79.5% on nearly balanced data.

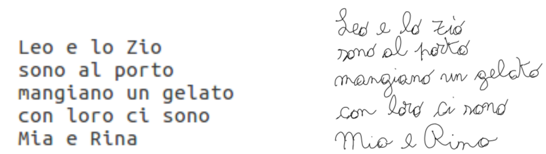


Figure 1: Italian BHK text and children writing example. English translation: "Leo and the uncle, are at the harbor, they are eating ice cream, with them are, Mia and Rina."

Contributions. In this study, the objective is to address the challenge of detecting dysgraphia in 9-10-years-old children. To achieve this, the participants have been asked to copy five Italian sentences. An algorithmic version of the BHK test was created, and a variety of characteristics were investigated to target both global

*Equal contribution. Names are displayed in alphabetic order.

DocEng '23, August 22–25, 2023, Limerick, Ireland

ACM ISBN 979-8-4007-0027-9/23/08.
https://doi.org/10.1145/3573128.3609351

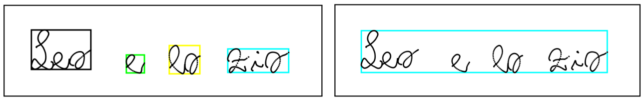


Figure 2: Words [left] and first row [right] after the validation phase.

and local writing information. To facilitate the research, a tablet and a smart pen were employed, using the *Bamboo Slate* device produced by Wacom[1]. This device allowed children to experience the same feeling of writing on a paper. and to record valuable information such as local and global spatial data, velocity, and more. To enhance the conducted analysis, a deep learning model has been used as a visual feature extractor. To validate the proposed approaches, two datasets comprising 95 samples from primary school children and 106 samples from adults were collected. The expertise of a pedagogist and the point of view of a group of primary school teachers have been taken into account as well, to provide a consistent labeling and to establish Human Level Performance (HLP) baselines. The main contributions of this work are as follows:

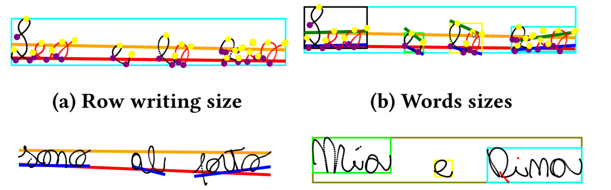
- **A new dataset for dysgraphia detection** in children handwritings, created within a reproducible setting using a commercial smart pen.
- **The use of a deep learning approach to extract visual features**, accompanied by a comprehensive discussion and comparative analysis of the results against standard assessment methods and HLP baselines.
- **The inclusion of teachers' perspectives in the evaluation process** as a crucial step for automatic screening. Working daily with children makes their point of view a valuable insights for timely dysgraphia intervention, helping to connect families and pedagogists.

2 METHOD

This section provides a detailed description of the developed algorithmic version of the BHK test (Section 2.1) and the trained deep learning architecture (Section 2.2). It also covers the pretraining techniques and image preprocessing methods employed in the study. The implementation of the methods described can be found in the GitHub[2] repository for reproducibility.

2.1 Algorithmic BHK

The data extraction process utilized the Bamboo Slate from Wacom, a screen-less tablet that emulates the pen and paper writing experience, ideal for young children. It captures writing input using a digital pen and stores the information in a dedicated app, allowing for easy export in various formats. The tablet records writing as strokes, represented by continuous paths of digital ink, with each stroke consisting of spatial coordinates, color, stroke thickness, and stroke order. The BHK test algorithm analyzes these strokes, grouping them into words and rows based on specific spatial thresholds and relative distances. To ensure accuracy, a validation phase confirms the expected number of words in a line. The Bamboo Slate's seamless integration of traditional writing with digital recording



(a) Row writing size **(b) Words sizes**

(c) Fluctuations **(d) Retouched or traced letters**

Figure 3: Various handwriting features extracted by the algorithmic BHK.

provides a practical solution for dysgraphia detection. By leveraging the captured strokes and employing the BHK test algorithm, meaningful word and row elements are obtained, as illustrated in Figure 2. This approach enables the accurate analysis of dysgraphia-related patterns and facilitates early identification and intervention for children with writing difficulties. As can be noticed in Figure 3, several parameters are selected from the handwriting for further analysis. As we can see in Figure 3, many parameters are selected from the writings for further analysis. The most salient features considered by BHK are: (Fig. 3a) *Row size*, for each word, the upper and lower local points are found, interpolated creating upper and lower lines and the average distances between these two lines takes into account; (Fig. 3b) *Word size, local maximum and minimum*, each word size is calculated by collecting words' local lower point [purple] local upper point [yellow]; (Fig. 3c) *Fluctuations*, as the relative slope between words and rows; (Fig. 3d) *Retouched or traced letters*, by analyzing temporal information of strokes and retouched elements [red]; and *Left margin alignment* calculated as the left slope margin of each line. The slope of the best-fit line indicates how much the left margin is aligned. This method involves gathering the set of features that the BHK test recognizes, resulting in a fixed-length feature vector with a total of 58 values, composed by 5 groups of features with 11 components each, corresponding to the five rows in the text, and 3 global features.

2.2 Deep-learning based approach

To assess the effectiveness of visual features alone in dysgraphia detection, in contrast to methods based on the BHK test, the ResNet18 architecture is employed, a compact variant of the ResNet architecture [8]. The obtained features were then used for classification. The results achieved through this approach are discussed in Section 4. The model has been pretrained using the IAM dataset [11], which consists of 657 writers and 13,353 isolated and labeled text lines, for the task of handwriting authorship identification. The model underwent a self-supervised training, employing a Triplet Loss to discriminate between different authors in the provided train, validation, and test splits. ResNet18, pretrained on ImageNet, was fine-tuned on handwriting data changing its head to match the binary task, necessary to align the model with the characteristics of dysgraphia, as demonstrated by the results. The new head uses two MLP, downsizing the hidden features from 512 to 100 and from 100 to "number of classes" dimensions, respectively, with LayerNorm and Dropout layers. Additionally, a separate dataset of adult handwriting samples has been collected to mitigate the limitations associated with children's handwriting, such as unbalanced data

[1]https://www.wacom.com/en-us/getting-started/bamboo-slate-and-bamboo-folio
[2]https://github.com/AILab-UniFI/dysgraphia-detection

distribution and dataset size. These considerations are further discussed in Section 3. The model underwent a second training phase using the collected adult data for a balanced binary classification task, specifically aimed at recognizing dominant and non-dominant hands. In all three data collections, the model learned at line level. The SVG files generated by the smart pen were processed to extract five lines from each collected handwriting sample. However, due to noise present in some SVG files, a few samples from both the adult and children data could not be accurately processed to produce their respective BHK features and segmentations. Consequently, these pages were excluded from the analysis to ensure the overall data quality, given the already limited sample size. The final dataset consisted of 92 out of 106 adult handwriting pages and 75 out of 95 children's pages. Each page was composed of five lines, which were split and then padded and resized to a predetermined dimension to retain crucial visual information. Subsequently, the pixel values of the lines were normalized. Every training used an AdamW optimizer (default parameters) and learning rate equal to $1e^{-5}$, with 32 samples per batch.

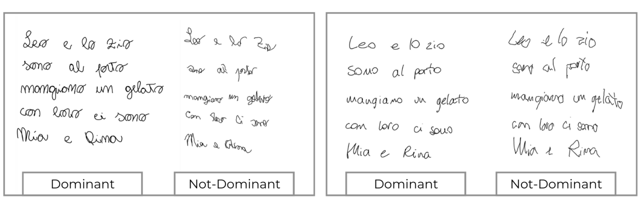
3 DATASET

To validate the proposed approach for dysgraphia detection, a dataset consisting of handwriting samples from 95 children has been gathered, resulting in a total of 475 lines. Additionally, some handwriting samples were collected from 53 adults, including both dominant and non-dominant hands, yielding 106 samples and 530 lines. The inclusion of adult handwriting aimed to bridge the gap between unsupervised pretraining on the IAM dataset [11] and the binary classification task for dysgraphia detection. The results are provided in Section 4. Participants were instructed to copy the BHK test sentences using the Bamboo slate, as shown in Fig. 4. For the children's handwriting samples, obtaining ground truth labels has been challenging. Certificates of dysgraphia were obtained for only a few children who visited an expert. Out of the total participants, only ten were officially diagnosed with dysgraphia. However, there was limited availability of "official" information for the remaining children, even though some exhibited symptoms of dysgraphia or other disorders, such as attention deficit, as also noted by their teachers. To address this issue and ensure a reliable and consistent dataset, an alternative labeling method was employed. To establish the ground truth for the children's handwriting, this work relied on the expertise of a pedagogist and a group of primary school teachers. The pedagogist (referred to as the EXPERT from now on) provided labels based on her opinion, regarding the presence or absence of dysgraphia in the samples. To preserve privacy, the children's handwriting samples were anonymized [3].

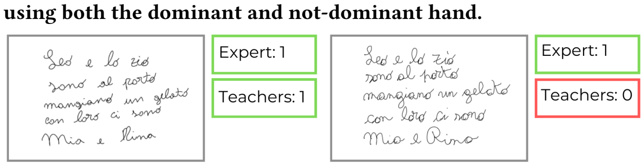
Furthermore, a panel of seven primary school teachers, selected for their experience with children and their ability to identify problematic handwriting, participated in the study. Each teacher independently evaluated the anonymized handwriting samples and responded to two questions:

- "Would you say the quality of this handwriting is below average (1), above average (0) or on average (2), based on your experience?"
- "Would you say this handwriting would require further investigations (1), it would not (0) or you are not sure (2), based on your experience?"

The teachers' responses generated a collection of information for each handwriting sample, resulting in 665 answers categorized into nine clusters. A majority voting approach was employed to assign each sample to a cluster, with the majority agreement among teachers determining the classification. As a result, 24 handwritings were classified as "generally good" ([0, 0]), 48 as "on average, not requiring further investigation" ([2, 0]), 10 as "on average but may require further investigation" ([2, 1]), and 13 as "below average and require further attention" ([1, 1]) (where [x, y] represents the answers to the first and second questions, respectively). This labeling process provides a human-level performance baseline, widely used in the literature, as exemplified in image classification by He et al. [9]. Using the second value of each cluster as a discriminator (0 or 1), the dataset comprised 72 handwritings classified as "generally good" or "not dysgraphic" and 23 handwritings classified as "may require further investigation" or "dysgraphic". Notably, the majority of samples (58 handwritings) were considered "on average", illustrating the challenge of distinguishing dysgraphia from non-dysgraphia classifications based solely on visual features.



(a) The two authors (gray boxes) have produced two handwritings using both the dominant and not-dominant hand.



(b) The image on the left show a mutual agreement between the EXPERT (pedagogist) and the (primary school) TEACHERS votes; on the contrary, the one on the right has been classified as a negative sample by the TEACHERS answers.

Figure 4: Adults (a) and children (b) dataset samples.

4 EXPERIMENTS

In this section, the baselines and performances of the proposed approach are presented, discussing the achievements and limitations of this work. A series of experiments have been conducted to explore the potentiality of using a deep learning architecture, such as ResNet18, in capturing important visual features for dysgraphia detection. Different training approaches are compared, including pretraining on the IAM dataset and a combined pretraining with the additional adult dataset. The performances of each approach was evaluated using precision, recall and F1, as detailed in Table 1: the results refer to a 4-fold cross validation on the proposed dataset

[3] Impossible to retrieve the child from the sample, showing nothing more than the handwriting only.

Table 1: Experimental results of dysgraphia detection on the 4-folds cross-validation setting. The "Pretraining" column indicates whether a pretraining stage was involved and, if so, with which dataset (IAM and adults principal/secondary hand). HLP stands for "Human Level Performances". The asterisk (*) indicates cases where *nan* values were obtained due to division by zero caused by the absence of true negatives and false negatives.

Method	Pretraining	Precision	Recall	F1
BHK + MLP	x	0.5*	0.5 ± 0.5	0.5*
Resnet18	x	0.5	1	0.67
Resnet18	IAM	0.5325 ± 0.019	**0.7707** ± 0.052	0.6055 ± 0.006
Resnet18	IAM + Adults	**0.726** ± 0.069	0.7 ± 0.074	**0.688** ± 0.086
HLP	x	0.819 ± 0.080	0.802 ± 0.063	0.800 ± 0.061

with different techniques and pretrainings involved. Firstly, to analyse the BHK features extracted and to pose an initial baseline, a two hidden-layers MLP classifier has been trained. As it is possible to see from Table 1, the results reveal that the MLP-based approach, with the BHK features, has shown limited discriminatory power, achieving a performance similar to random guessing. This could have been due to the challenges in distinguishing between dysgraphic and non-dysgraphic samples and / or to the limited set of features extracted with the proposed BHK algorithm. As a second step, ResNet18 has been employed to discover how much could have been achieved with a bigger feature space relying only on visual artifacts. The version without any pretraining, e.g. ResNet18 with ImageNet weights, could not detecting anything but "dysgraphic", indicating the need for more informative features. To further enhancing these preliminary results, two more training phases have been employed. The ResNet18 pretrained on the IAM dataset showed an important improvement in performances, since the predictions do not collapse to one unique class as it can also be noted by the precision and recall values. Thanks to an additional pretraining on the adults dataset, the model could achieve better results, yielding the best F1 score among the different experiments and finding a good balance between precision and recall. Despite the higher F1 that reduce the gap to HLP of about 8%, a drawback of 7% could be noticed from the Recall score compared to only IAM pre-training, probably due to noise introduced by adults handwritings with no SLD. The HLP scores were evaluated using the interviews' results as *"predictions"*, using the same validation splits: there is still an important margin to be closed within the best model obtained and the teachers performances. Future works in this direction are discussed in the next section.

Limitations and Future Works. There are some limitations to this research that could be used as opportunities for future work:

- *the limited dataset size*: gathering more and balanced samples is not an easy task, but it would drastically help for better results and generalization purposes making multi-lingual handwritings an opportunity to better prove the importance of relying only on visual features;
- *using only ResNet18*: the choice of a small convolutional network has been motivated also by the small dataset. Having a larger dataset would enable to try other architectures to compare with, such as visual transformers;
- *only adults data as augmentation*: exploring more augmentation techniques suitable for dysgraphia would be crucial, helping to retain important recall performances for the ultimate purpose of screening.

5 CONCLUSIONS

In this work the use of a smart-pen and a visual extractor for early identification of dysgraphia in children has been described. An algorithmic versions of the BHK test has been implemented and a deep-learning architecture was used to tackle the downstream task. Two datasets of children and adults handwriting were collected for the experiments, taking into consideration a pedagogist and teachers expertise. Despite the limited number of samples and the class unbalance, the proposed pretraining let the model achieve good results. Finally, limitations of this work have been discussed, proposing future directions to which this research could be extended.

REFERENCES

[1] Esther Adi-Japha, Yael E Landau, Lior Frenkel, Mina Teicher, Varda Gross-Tsur, and Ruth S Shalev. 2007. ADHD and dysgraphia: underlying mechanisms. *Cortex* 43, 6 (Aug. 2007), 700–709.

[2] Thibault Asselborn et al. 2018. Automated human-level diagnosis of dysgraphia using a consumer tablet. *NPJ Digit. Med.* 1, 1 (Aug. 2018), 42.

[3] Maëlle Biotteau et al. 2019. Developmental coordination disorder and dysgraphia: signs and symptoms, diagnosis, and rehabilitation. *Neuropsychiatr. Dis. Treat.* 15 (July 2019), 1873–1885.

[4] Giovanni Dimauro, Vitoantonio Bevilacqua, Lucio Colizzi, and Davide Di Pierro. 2020. TestGraphia, a Software System for the Early Diagnosis of Dysgraphia. *IEEE Access* 8 (2020), 19564–19575. https://doi.org/10.1109/ACCESS.2020.2968367

[5] Peter Drotár and Marek Dobeš. 2020. Dysgraphia detection through machine learning. *Scientific reports* 10, 1 (2020), 21541.

[6] Katya P Feder and Annette Majnemer. 2007. Handwriting development, competency, and intervention. *Dev. Med. Child Neurol.* 49, 4 (April 2007), 312–317.

[7] Thomas Gargot et al. 2020. Acquisition of handwriting in children with and without dysgraphia: A computational approach. *PLoS One* 15, 9 (2020), e0237575.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

[10] Ragnheidur Karlsdottir and Thorarinn Stefansson. 2002. Problems in developing functional handwriting. *Percept. Mot. Skills* 94, 2 (April 2002), 623–662.

[11] U-V Marti and Horst Bunke. 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition* 5 (2002), 39–46.

[12] Chung PJ, Patel DR, and Nizami I. 2020. Disorder of written expression and dysgraphia: definition, diagnosis, and management. *Transl Pediatr. ;9(Suppl 1):S46-S54.* doi:10.21037/tp.2019.11.01 (2020).

[13] Giulia Silveri et al. 2019. Influence of Dysgraphia on Kinematic Characteristics of Handwriting in Italian Primary School Children. In *World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 2).* Springer, 241–245.

[14] B C Smits-Engelsman, A S Niemeijer, and G P van Galen. 2001. Fine motor deficiencies in children diagnosed as DCD based on poor grapho-motor ability. *Hum. Mov. Sci.* 20, 1-2 (March 2001), 161–182.

[15] Hilde Van Waelvelde, Tinneke Hellinckx, Wim Peersman, and Bouwien CM Smits-Engelsman. 2012. SOS: a screening instrument to identify children with handwriting impairments. *Physical & occupational therapy in pediatrics* 32, 3 (2012), 306–319.